RDF Validation in the Cultural Heritage Community

Introduction

Tutorial @ Dublin Core 2014
October 8th, 2014
Austin, TX, USA
Slides and Materials

Your friendly lecturers
(in order of appearance)

• **Kai Eckert**  
  *University of Mannheim, Germany*

• **Stefanie Rühle**  
  *State and University Library (SUB) Göttingen, Germany*

• **Karen Coyle**  
  *Consultant, USA*

• **Tom Baker**  
  *Sungkyunkwan University, Korea*

• **Thomas Bosch**  
  *GESIS – Leibniz Institute for the Social Sciences, Germany*

• **Tom Johnson**  
  *Digital Public Library of America (DPLA), USA*
Schedule

- 1:00 - 1:30  Introduction, Validation and Application Profiles in a nutshell
  Kai Eckert, Stefanie Rühle

- 1:30-2:00  RDF and Validation: What is the issue?
  Karen Coyle, Tom Baker
  » Difference between inferencing and validation
  » OWL in an open world

- 2:00-3:00  Validation techniques in use
  Thomas Bosch
  » Practical examples, Demonstation

- 3:00-3:30  BREAK

- 3:30 - 4:15  Validation on current data models
  Stefanie Rühle, Tom Johnson
  » Europeana Data Model
  » Digital Public Library of American Data Model

- 4:15 - 5:00  Application profiles as a possible solution
  Karen Coyle, Tom Baker
  » Dublin Core Application Profiles and the Singapore Framework
  » BIBFRAME profiles
  » DCMI RDF validation group work
Learning Objective

• Attendees will learn why validation is an issue with RDF and OWL, and how constraints in those languages differ from our usual concept of constraints.
• Attendees will learn what efforts exist today to solve this problem.
• Attendees will get a brief overview of some particular validation issues with cultural heritage data.
• In particular, attendees will be able to ask questions of speakers with "real life" experience with linked data and validation.
Metadata Provider vs. Open World

Open World:
„Anyone can say anything about anything.“

Just don’t disagree!
That’s somehow not working for us...

We want to know who said what. And how. And if it is correct. And by the way...

... where would this lead us...
How everything started: SWIB 2013

... how to express constraints on ones own data,
... how to validate data according to these constraints,
... how to provide data in different flavors (DC, Bibframe, BibO, ... → DM2E, EDM, ...)
... how to access data on a server in a certain flavor.
The problem with constraints

Example: RDFS Domain and Range

dcterms:creator rdfs:range foaf:Agent .

This is not really a constraint, but a statement:
„Whatever you name as creator, it is a foaf:Agent.“
Recently in the hammer business

\[
:me \text{ ex:hasProblem :problem1 .} \\
\text{ex:hasProblem rdfs:range :nail.}
\]

I don’t constrain my problems to be nails. I simply state: If I have a problem, then it is a nail.

\[
:me \text{ ex:hasProblem :badAtExamples .} \\
\Rightarrow \text{:badAtExamples a :nail . (q.e.d.)}
\]
Open World Pollution

dcterms:creator rdfs:range ex:PrimaryAuthor .

This is possibly correct for my (local) application.

Globally, this statement might lead to conflicts (within the data).
Typical approach

We create our own vocabulary:

```
ex:hasPrimaryAuthor
   rdfs:subPropertyOf dcterms:creator.
```

Semantically everything is sound now.
Linked Data Practice

Mix and match, reuse vocabularies.

- No reasoning required.
- Data is better understandable (?).

Own vocabulary vs. reuse... sounds familiar?
This debate is as old as Linked Data (at least...)

2014-06-12
What we really want...

For our data, we want to state,

... how values are represented exactly,

... which values are mandatory,

... which values are mutually exclusive,

... and much more (cataloguing rules?).

Open World means: Everybody is free to follow our rules. At least we follow them and you can rely on it.
Application profiles for RDF

• Mix-and-match of existing vocabularies.
• Local constraints on the data.
• Machine-processable and accessible together with the data.
• Support of different Application Profiles by client and/or server possible.

So far our rough idea...
RDF Application Profiles Task Group

• Begin: June 10, 2014
• Duration: 1 year
• Chair/Co-Chair: Antoine Isaac and Karen Coyle
• Liaison with technical board and community spec committee: Kai Eckert and Valentine Charles
• Editorial board: Evelyn Dröge and Thomas Bosch
• Form: DCMI Task Group, reporting to the Community Specifications Committee
• Wiki: http://wiki.dublincore.org/index.php/RDF_Application_Profiles/
R DFA-AP DCMI Task Group: People

Kai Eckert, University of Mannheim, DM2E, Germany
Evelyn Dröge, HU Berlin, DM2E, Germany
Steffen Hennicke, HU Berlin, DM2E
Julia Iwanowa, HU Berlin, DM2E, Germany
Konstantin Baierer, HU Berlin, DM2E, Germany
Doron Goldfarb, Austrian National Library, DM2E, Austria
Stefanie Rühle, SUB Göttingen, Germany
Tom Baker, DCMI, USA
Martin Malmsten, Swedish National Library, Sweden
Niklas Lindström, Swedish National Library, Sweden
Rurik Greenall, NTNU Library, Norway
Lars G. Svensson, German National Library, Germany
Jan Polowinski, TU Dresden, Germany
Antoine Isaac, Europeana, Netherlands
Valentine Charles, Europeana, Netherlands
Robina Clayphan, Europeana, Netherlands

Georgios Markakis, Europeana, Netherlands
Karen Coyle, USA
Bernard Vatant, Mondeca, France
Gordon Dunsire, IFLA/JSC, Scotland
Adrian Pohl, hbz, Germany
Mark Matienzo, Digital Public Library of America, USA
Corey Harper, New York University, USA
Mariana Curado Malta, University of Minho, Portugal
Thomas Bosch, Gesis - Leibniz-Institute for the Social Sciences, Germany
Miika Alonen, CSC - IT Center for Science, Finland
Diane Hillmann, Metadata Management Associates, LLC
Dickson Lukose, MIMOS Berhad, Malaysia
Matthias Palmér, MetaSolutions, Sweden
Nor Azlinayati Abdul Manaf, MIMOS Berhad, Malaysia
Case Studies / Use Cases

1. DPLA RDF application profile use cases
2. Digitised Manuscripts to Europeana (DM2E)
3. Europeana Data Model
4. DINI AG KIM - RDF-Representation of Bibliographic Data
5. Use of EDM in the Deutsche Digitale Bibliothek
6. OER World Map prototype
7. Reusing CEN EuroLMAI, DC etc. in academia
8. Using RFC 6906 Profiles to specify a profile independently of the media type
Tasks

The RDF Application Profile Task Group will:

1. **gather experts from theory and practice** dealing with the problem areas sketched in the introduction,
2. collect and describe case studies from these experts and the general public,
3. extract common use cases from these case studies that illustrate particular problems,
4. specify requirements to be fulfilled in order to adequately solve these problems and meet the use cases,
5. investigate existing best-practices regarding these requirements,
6. identify gaps and **recommend best-practices** to close them,
7. test the proposed solution on selected relevant profiles.
DM2E Use Case

• The DM2E model *specialises the EDM* by...
  … offering EDM properties needed to model manuscripts
  … adding new classes and properties from other vocabularies
  … adding new properties in the DM2E namespace
• Resources are mixed and matched
• DM2E-specific definitions and instructions via property `dm2e:scopeNote`
• DM2E-specific definitions are not violating original ones
Element definitions in DM2E

• Different definitions
  – Original definition: Often broader
  – DM2E scope note: How to use the element in our specific context
  – Example: Definitions of dc:creator

<table>
<thead>
<tr>
<th>Dublin Core</th>
<th>EDM</th>
<th>DM2E</th>
</tr>
</thead>
<tbody>
<tr>
<td>An entity primarily responsible for making the resource.</td>
<td>An entity primarily responsible for making the resource. This may be a person, organisation or a service.</td>
<td>The property dcterms:creator holds the name or identifier of the agent (a person or organisation) who created the aggregation, i.e. the original metadata record, or CHO (possibly its author).</td>
</tr>
<tr>
<td></td>
<td>For the creator of the CHO. If possible supply the identifier of the creator from an authority source. Repeat for multiple creators.</td>
<td></td>
</tr>
</tbody>
</table>

• Data must be valid according to all definitions
Validation of DM2E data

• Validation checks before ingestion:
  – Cardinality constraints
    • E.g. check if mandatory elements are delivered
    • Example: `dm2e:setDisplayLevel` must be mapped exactly once in `ore:Aggregation`
  – Complex interrelated constraints
    • Example: The value of `edm:end` must not be before `edm:begin`
  – Exclusively usage of elements from the DM2E model
    • Check whether a property is part of the model
    • Example: `dcterms:creator` is not part of the model and cannot be used whereas `dc:creator` is allowed
  – URL syntax
    • E.g. check path hierarchy and allowed characters
    • Example: `http://www.example.de/tüdelü` is not allowed
Validation of DM2E data II

• Checks (continued):
  – Literals
    • E.g. check datatypes
    • Example: Check whether xsd:datetime (recommended for dcterms:created and similar properties) is correctly used
  – Domain and range
    • Literal and resource usage (1), correct domains and ranges (2)
    • Example 1 (no Literal allowed):
      dm2edata:item/uber/dingler/ar00101
dc:publisher “Dingler”@de.
    • Example 2 (no Agent allowed):
      dm2edata:item/uib/wittgenstein/Ms-114
dcterms:references
dm2edata:agent/uib/authority_gnd/118634313.
  – Enforce namespace changes
    • E.g. use the correct DM2E namespace
    • Example: http://onto.dm2e.eu/schemas/dm2e/0.1/ is not allowed
Validation rules

- Validation of DM2E data
  - Automated validation per RDF file before ingestion
  - The validator gives feedback to the provider

<table>
<thead>
<tr>
<th>Feedback groups</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOTICE</td>
<td>for things that are not a problem but may have eluded the mapper's attention</td>
</tr>
<tr>
<td>WARNING</td>
<td>for things that are not errors, but that contradict strong recommendations in the model</td>
</tr>
<tr>
<td>ERROR</td>
<td>for things that contradict strict assertions in the model, such as mandatory elements missing</td>
</tr>
<tr>
<td>FATAL</td>
<td>for errors that cripple the data to an extent where it can't be worked with sensibly</td>
</tr>
</tbody>
</table>
Validation challenges

• Problem: Automatic validation based on the model
  – No machine-readable representation of constraints
  – Against which level of specialisation should be validated?
    • All definitions must be fulfilled!
  – How to deal with contradictions?

• Common misunderstanding:

Reasoning ≠ Validation
SPIN, SPARQL, OWL, DSP...

```
SELECT ?this ?subope ?object WHERE {
  ?C owl:allValuesFrom :Dog .
  ?C a owl:Restriction .
  ?this rdf:type ?subC .
  ?subC rdfs:subClassOf* ?C .
  ?this ?subOPE ?object .
  ?subOPE rdfs:subPropertyOf* :hasPet .
  FILTER NOT EXISTS { ?object rdf:type :Dog . }
}
```

```
<StatementTemplate type="nonliteral">
  <Property>:hasPet</Property>
  <NonLiteralConstraint>
    <ValueClass>:Dog</ValueClass>
  </NonLiteralConstraint>
</StatementTemplate>
```
Schedule

• 1:00 - 1:30  Introduction, Validation and Application Profiles in a nutshell
  *Kai Eckert, Stefanie Rühle*

• 1:30-2:00  RDF and Validation: What is the issue?
  *Karen Coyle, Tom Baker*
  » Difference between inferencing and validation
  » OWL in an open world

• 2:00-3:00  Validation techniques in use
  *Thomas Bosch*
  » Practical examples, Demonstation

• 3:00-3:30    BREAK

• 3:30 - 4:15  Validation on current data models
  *Stefanie Rühle, Tom Johnson*
  » Europeana Data Model
  » Digital Public Library of American Data Model

• 4:15 - 5:00 Application profiles as a possible solution
  *Karen Coyle, Tom Baker*
  » Dublin Core Application Profiles and the Singapore Framework
  » BIBFRAME profiles
  » DCMI RDF validation group work